教育信息数据挖掘初探

黄　成

（海南师范大学　海南　海口　571158）

【摘　要】随着人类社会跨入信息时代，以计算机技术和网络技术为核心的信息技术正对教育产生着深刻的影响，并成为教育者作出教育决策的主要技术基础和手段。准确有效的教育信息是教育者作出正确决策的主要依据，信息技术的发展给教育带来了大量的教育信息数据，本文就教育信息数据挖掘的概况以及如何在大型的教育数据库的数据中挖掘出有用的教育决策信息做初步的探讨。

【关键词】数据挖掘；关系数据库；模式；聚类分析；教育信息
【中图分类号】G434　　【文献标识码】C　　【文章编号】1001－8700（2006）04－0064－03

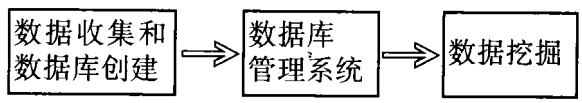
随着人类社会跨入信息时代，信息技术正在对教育产生着深刻的影响，并成为教育改革的技术基础和强大动力。在教育领域全面深入地运用现代信息技术来促进教育改革和教育发展过程，其结果必然是形成一种全新的教育形态信息化的教育。从教育层面上看，信息化教育具有教材多媒化、资源全球化、教学个性化、学习自主化、活动合作化、管理自动化、环境虚拟化和系统开放化等特点。教育的决策者（包括教育部门领导、教师、学习者等）可以从丰富的教育信息中作出自己的教育决策，从而提高教育决策者的工作或学习效率，避免出现工作或学习的偏差或错误。然而随着计算机和网络技术的快速发展，快速增长的海量教育信息数据收集、存放在大型的数据库中，理解它们已远远超出人的能力，结果大量的数据变为无用的垃圾，从而造成重要的教育决策往往来自于教育决策者的直觉。这种情况被称为教育数据丰富，但教育信息贫乏。

一、信息时代教育信息数据挖掘的必要性

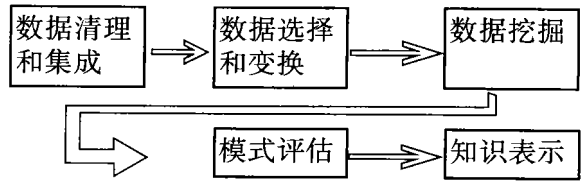
教育信息化是实现素质教育的重要步骤，是教育现代化的基本内容。信息化包含有两种含义，其一是对教育信息重要程度的认识，应将信息作为系统中的一种基本要素进行处理。在教育信息化的过程中，首先应对系统中的信息和信息的作用进行分析；其二是在信息分析的基础上将信息真正地为教育服务。数据挖掘之所以应引起教育界的重视是因为大量存放的教育数据可以广泛使用，并且迫切需要将这些教育信息数据转换成为有用的信息和知识。然而由于缺乏强有力的数据分析工具，结果收集的在大型数据库中的数据变成了难得再访问的数据档案。这样教育决策者重要的决定常常不是基于数据库中的有用信息，而是基于决策者的直觉。这就不可能实现信息在教育领域中的有效应用，也就不可能真正实现有效的教育信息化。

二、数据挖掘和教育信息数据挖掘

数据挖掘就是从大量的数据中提取或挖掘有用的信息或知识。获取的信息和知识可以广泛应用，包括商务管理、生产控制、市场分析、工程设计和科学探索等。数据挖掘是信息技术自然发展的结果，其演变的情况如下图：



数据收集和数据库创建机制的早期开发已成为稍后数据存储和检索、查询和事务处理有效机制开发的必备基础。随着提供查询处理和事务管理的大量数据库系统广泛付诸实践、数据分析和理解自然成为数据处理的一个目标。数据挖掘又常被称为数据库中的知识发现，它的执行步骤一般如下图：



作者简介：黄成，海南师范大学讲师，硕士。

数据清理和集成主要是清除不一致的数据并把多种数据源组合在一起；数据选择和变换从数据库中检索和分析与目的相一致的数据并进行变换成为适合挖掘的数据形式。数据挖掘使用智能方法提取数据模式，根据某种兴趣度度量，识别表示知识的真正有趣的模式。然后使用可视化和知识表示技术向用户提供控制的知识。可见，真正的数据挖掘只是知识发现的关键一步。

随着我国教育信息化进程的不断深入，数据库建设已成为信息化教育的重要核心之一。教育数据库是指为了满足教育需要的，为教育管理者提供教育管理和决策支持的相关信息；为教师提供丰富的备课用参考资料；为学生提供海量的 CAI 软件以满足学科齐全、内容丰富和形式多样的教学要求，以适合不同年级、不同学科课堂教学的需要和不同能力的个别化学习者需要的数据集合。一个大型的教育数据库通常包含教育中各方面的相关信息，可以说是包罗万象，数据量相当惊人。而通过网络又可以把许多这样的数据库连接起来，构成一个庞大的网络化教育信息资源。这些大的数据量会不会出现数据坟墓呢？如果没有强有力的教育数据分析工具，答案显然是肯定的。教育数据和智育信息之间的鸿沟要求人们系统地开发数据挖掘工具，将难于理解的教育数据转换成有用的教育信息和知识。

三、教育信息数据挖掘的模式

教育信息数据存储在各类的数据库中，如关系数据库、数据仓库、展开文件和 WWW 等，原则上讲，数据挖掘可以在以上任何一类的信息存储上进行。数据挖掘系统必须在存储系统上挖掘出多种类型的模式以适应不同教育用户的要求，教育用户清楚自己的教育信息需求，然而他们不知道数据中有什么类型的模式是有趣的，因此一般用户都想进行并行地搜索多种不同的模式。此外，教育数据挖掘系统应能够发现不同的抽象层次的模式，能够允许用户给出提示、指导或聚焦用户有趣的模式的搜索。同时有些模式并非对数据库中所有的数据都成立，通常每个被发现的模式带上一个确定性或可信性度量。教育信息挖掘可以发现的模式类型一般有以下几种：

（一）数据特征化和数据区分

数据特征化是目标类数据的一般特征或特性的汇总而数据区分是将目标类对象的一般特性与一个或多个对比类对外的一般特性比较。在教育信息中，用汇总的、简洁的、精确的方式描述每个类和概念是十分有用的。这种类或概念的描述称为类/概念描述。

在远距离的网络课程教学中，教学过程数据和教学评价数据可通过网络关系数据库进行收集。对于教学决策者来说，教学过程反馈数据和学生评价数据的特征化和区分具有重要意义。因为通过它教育决策者可以了解教学的成效程度，以便对课程教学进行调控。通常教育决策者可通过数据库查询他们所感兴趣的模式。例如，经过一段时期的教学，教师可能想知道哪些学生经常光顾课程教学网站，他们的学习活跃程度如何，是否经常参与教学交流，他们的学习成绩与这些因素存在什么关系等。结果可能发现学习成绩好的学习者一般轮廓，如学习基础较好、经常参与网络课程学习，善于与教师和学习伙伴进行交流。教育者也可能通过教育信息挖掘系统进行统计上网学习的次数与学生学习成绩的关系来发现更多的学生者的特征。数据特征和区分的输出可以采用多种形式提供不同用户的需求，如拼图、条形图、曲线和多维表等。

（二）关联分析

关联分析是指发现数据项集之间的相关联系。随着数据的不断收集和存储，在数据库中收集的项目越来越多，这些项目之间有何种关联呢？这也许是我们所感兴趣的。

在传统的教学模式中，教育工作者通过各类测试收集学生各门科目的考试成绩。试图从各种科目成绩中研究不同学科的相关性来了解不同学科之间的关系。这在学科教学研究中是必要的，因为它可以为学科的课程设置带来帮助。传统的作法是通过收集多门课程的成绩样本利用传统测试数据的统计方法计算相关系数。显然传统的作法既费时效率又不高。随着教育数据库的发展，这些数据都可以存储在数据库中，通过数据的关联挖掘可以得到相关的信息。同样在教育信息的结构分析中，学生—问题表（即 S—P 表）中对于问题和学生得分的关联性分析可为学习的诊断、教学的评价提供十分重要的信息。通过教育信息数据关联挖掘显然能为我们提供十分重要的帮助。

（三）分类和预测

分类是指找出描述并区分数据类或概念的模型（或函数）以便能够使用模型预测类标记未知的对象类。分类可以用来预测数据对象的类标记，然而在某些应用中，人们可能希望预测某些空缺的或不知道的数据值而不是类标记，当被预测的值是数值数据时，通常称之为预测。预测可能涉及数据值预测和类标记预测，通常预测限于值的预测，并因此不同于分类。预测也包含基于可用数据的分布趋势

识别。

在教育系统中，对教育对象进行分类和预测是必要的。例如，对学生的成绩分类、对学生能力分类、对教师的教学水平分类等等。对于每一位学生的数据，对于每一位教师的数据，应如何决定其所属的类别，这是一种分类的问题。分类分析（教育中常称为判别分析）是教育经典统计方法之一。同样在教学过程中，对于学生以往的数据预测学生学习的趋势能为教师进行调控教学提供重要的教学信息。尤其在网络教学中，教学主要是通过网络课程来进行，对学生学习数据进行前期的预测显得尤为重要。

（四）聚类分析

聚类分析数据对象，而不考虑已知的类标记。聚类可以用于产生类标记，它主要根据最大化类内的相似性、最小化类间的相似性的原则进行聚类。聚类分析可使用设定的距离范围，在范围之内为同一类，超出范围为不同类。在许多应用中，可以将一个类中的数据对象作为一个整体来对待并加以分析应用。

在机器学习领域，聚类分析在无指导学习中具有广泛的应用。同样在教育测试和评价中，如何根据测试的结果对学生的成绩进行聚类分析能为我们提供相关的教育信息。

（五）孤立点分析

数据库中可能包含一些数据对象，它们与数据的一般行为或模型不一致。这些数据对象是孤立点。大部分数据挖掘方法将孤立点视为噪声或异常而丢弃。然而，在一些应用中，罕见的事件可能是我们值得关注的。孤立点探测和分析是一个有趣的数据挖掘任务，被称为孤立点挖掘。孤立点可以使用统计试验检测，它假定一个数据分布或概率模型，并使用距离度量，到其他聚类的距离很大的对象被视为孤立点。基于偏差的方法通过考察一群对象主要特征上的差别识别孤立点，而不是使用统计或距离度量。

教学的过程性评价、总结性评价和学生的学力评价在教育中具有十分重要的作用，评价指标能了解和甄别学生掌握该阶段教学目标的情况，区分学习水平，评定学习成绩优劣。学力评价是为选拔服务而进行的，而过程性评价和阶段性总结性评价主要为教学提供反馈调节、诊断指导、强化激励的功能。因而孤立点分析在日常教学过程中是必要的。同样，在教育信息分析过程中，对于孤立点的分析可以让我们清楚教育过程中的异常情况，从而采取有效的补救措施。

（六）演变分析

数据演变分析是指描述行为随时间变化的对象的规律或趋势，并对其建模。尽管它可能包括时间相关数据的特征化、区分、关联、分类或聚类，这类分析的不同特点包括时间序列数据分析、序列或周期模式匹配和基于类似性的数据分析。

四、教育信息数据挖掘的应用与意义

数据挖掘是一种知识发现的技术，它能从关系数据库、数据仓库、展开文件和 WWW 等各种数据集合中挖掘出用户所需要的知识，真正作到将数据坟墓转换成为“金块”。在信息化社会的今天，这一技术的应用与挖掘显得尤为重要。

数据挖掘是一个新兴的领域，目前主要应用于金融、医学、电信等行业。尽管如此，每年市场上都会出现新的数据挖掘系统，而且不断增加新的功能和特性以适应用户的需求。

信息化教育是我国教育发展的未来趋势，重视教学过程的信息分析是实现教育信息化的基础和条件。因此，信息化教育应研究如何基于信息科学对教学系统中的各种信息进行处理，并将这些处理的结果有效地用于完善教学系统的设计、控制和评价中。然而随着计算机技术和网络技术的发展，教育信息数据的丰富一方面为我们教育/教学提供有用的教育信息，另一方面教育信息数据的不断膨胀导致我们面对大量的数据不知何从，加上数据噪声的干扰进一步扰乱我们的视线。大量数据变为难于再访问的数据档案，教育信息化也难于真正地、有效地进行。教育信息数据挖掘技术在教育中的广泛应将对我国教育信息化的进程起着重要的推进和支撑作用。

【参考文献】

[1] Jiawei Han Micheline Kamber 著范明、孟小峰等译，数据挖掘概念与技术，机械工业出版社，2005.

[2] 傅德荣，章慧敏，教育信息处理，北京师范出版社，2003.

[3] 祝智庭，钟志贤，现代教育技术，华东师范大学出版社，2003.

[4] Abraham Silberschatz Henry E. Korth S. Sudarshan 著，杨冬青，唐世渭等译，数据库系统概念，电子工业出版社，2005.

[5] 罗黎辉、高翔，教育测量与评价，云南教育出版社，2002.

（本文责任编辑：陈　新）